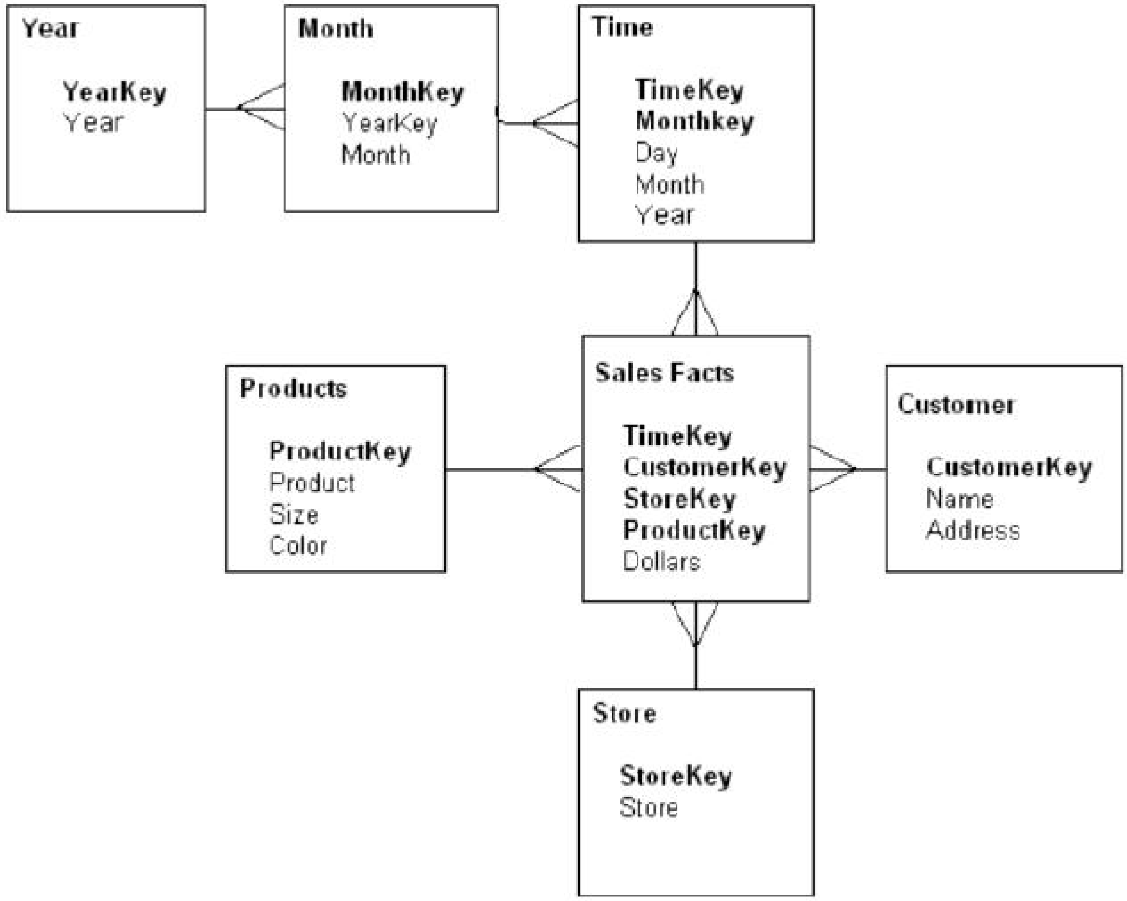
DatawareHouse Assessmet

1. For the given Dimensional Modelling, please identify the following:



1.1 How many dimensions and Facts are present?

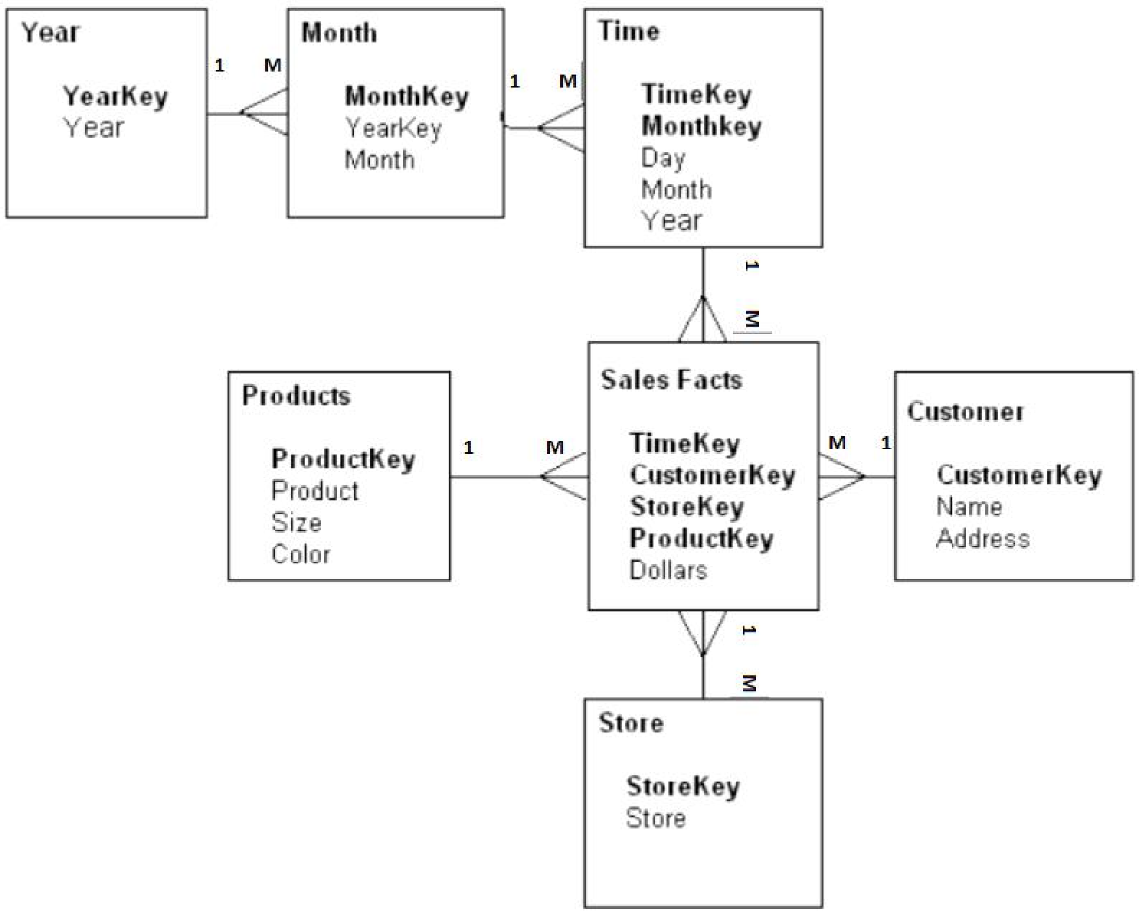
There are 6 Dimension and 1 Fact table in the given Dimensional Model.

- Fact Table : Sales Fact.
- Dimension Table: Store, Customer, Product, Time, Month, Year.

Month and Year are the Normalized Dimension of the Time Dimension Table.

1.2 Please identify the cardinality between each table?

Cardinality of the table is,



Customer : SalesFact ==> 1 : M

Store : SalesFact ==> 1 : M

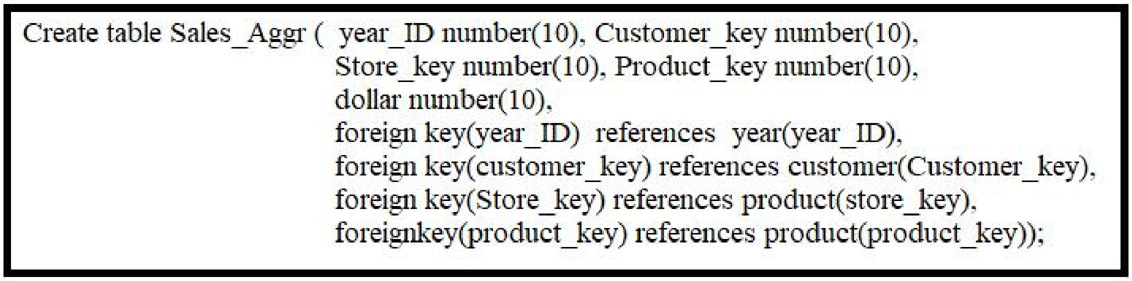
Product : SalesFact ==> 1 : M

Time : SalesFact ==> 1 : M

Month : Time ==> 1 : M

Year : Time ==> 1 : M

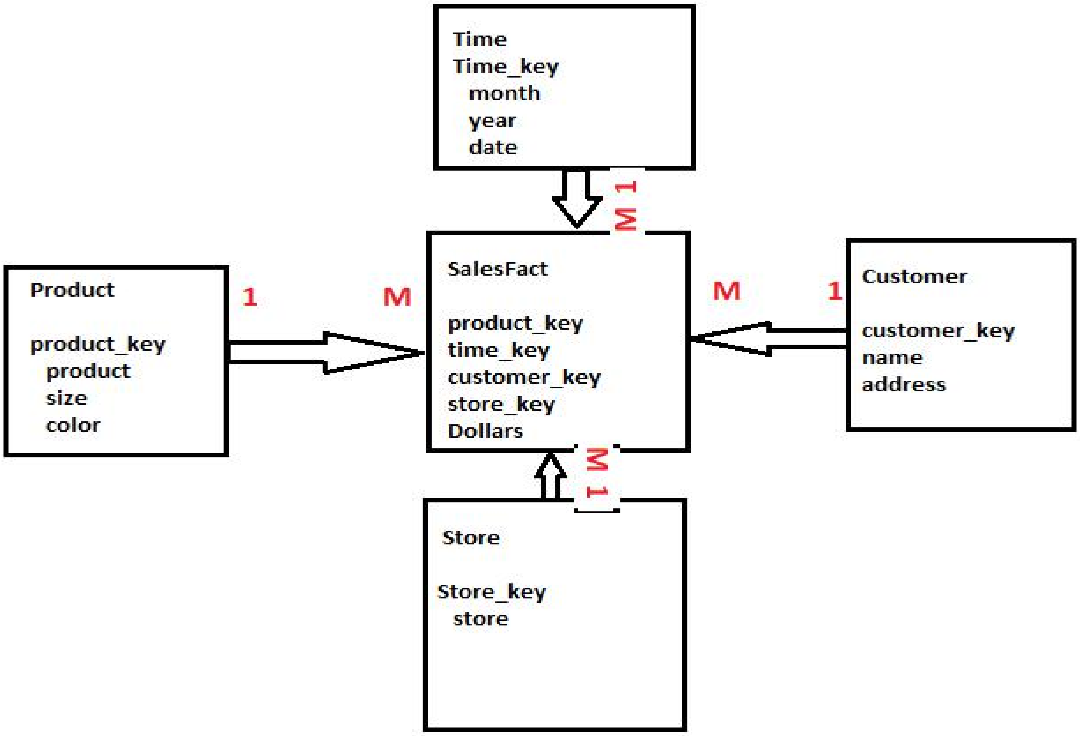
1.3 How to create a Sales_Aggr fact using the following structure (SQL Statement):



```
Create table Sales_Aggr (  year_ID number(10), Customer_key number(10),
                           Store_key number(10), Product_key number(10),
                           dollar number(10),
                           foreign key(year_ID)  references  year(year_ID),
                           foreign key(customer_key) references customer(Customer_key),
                           foreign key(Store_key) references product(store_key),
                           foreignkey(product_key) references product(product_key));
```

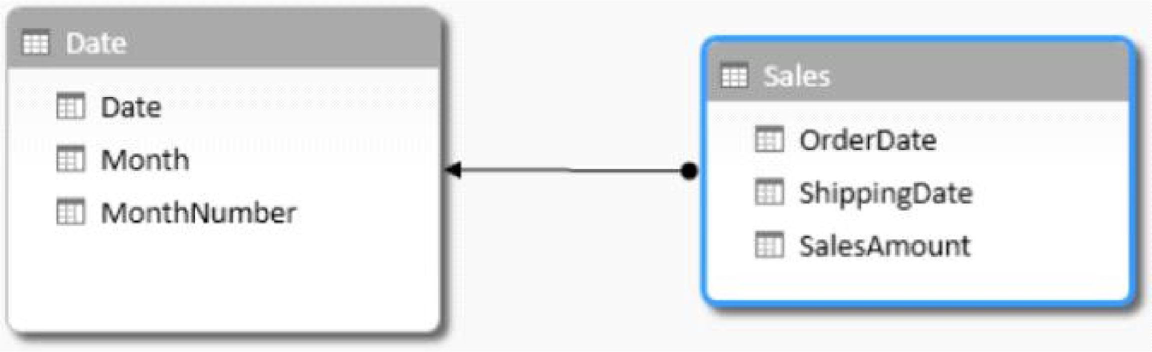
1.4 Can you Please Modify the above snowflake schema to Star schema and draw the dimension model, showing all the cardinality?

The Star Scheme For the given Snowflakes is,

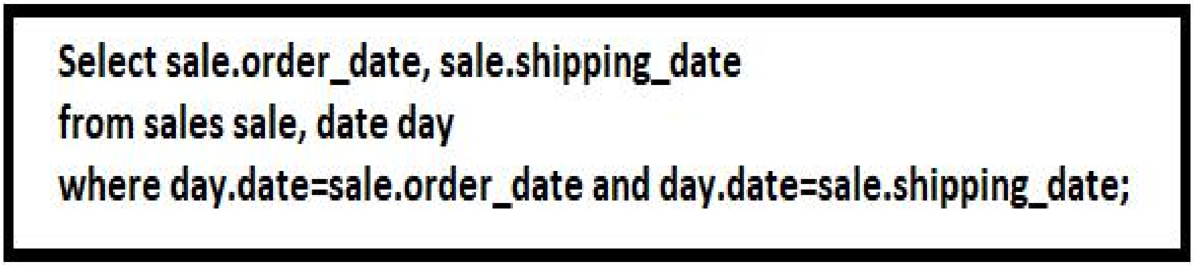


==

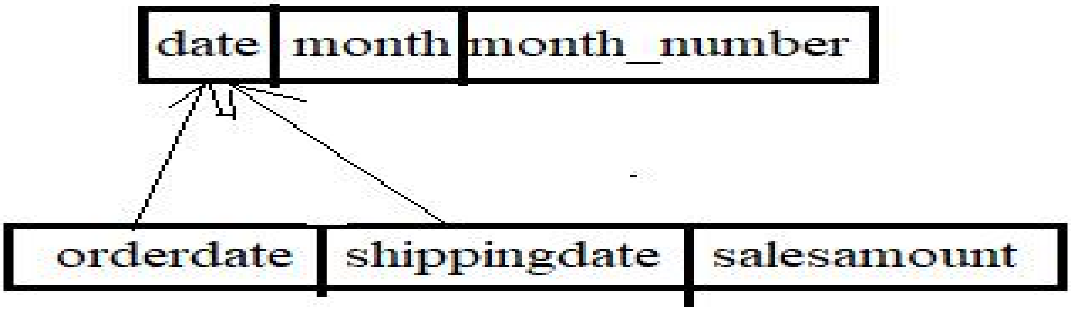
2. For the following dimension Model can you please give an example of Circular Join and how to avoid it:



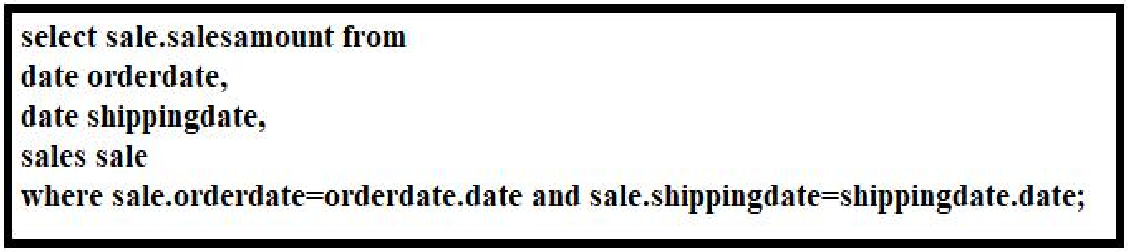
In the given Schema the Sales table has 2 dates i.e., order_date and shipping_date. This date are different most of the time but the Date dimension has only one date Attribute, which will be pointing to the both order_date and shipping date. This type of reference of two different date to a single attribute is called circular join. Below is the example of circular join.



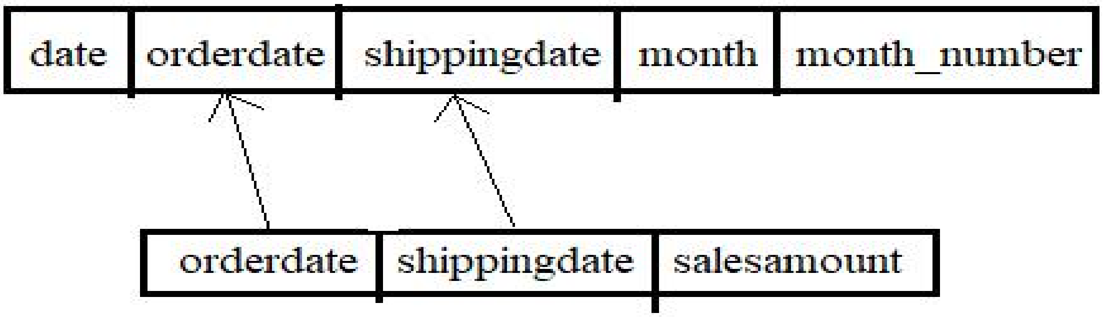
```
Select sale.order_date, sale.shipping_date
from sales sale, date day
where day.date=sale.order_date and day.date=sale.shipping_date;
```



To overcome the circular join, we can use the new attributes in the date dimension which will have date with the alias name as orderdate and shippingdate through which the sale table orderdate and shippingdate can be mapped to the date table orderdate and shippingdate. hence overcome the circular join. The query for non-circluar join is,

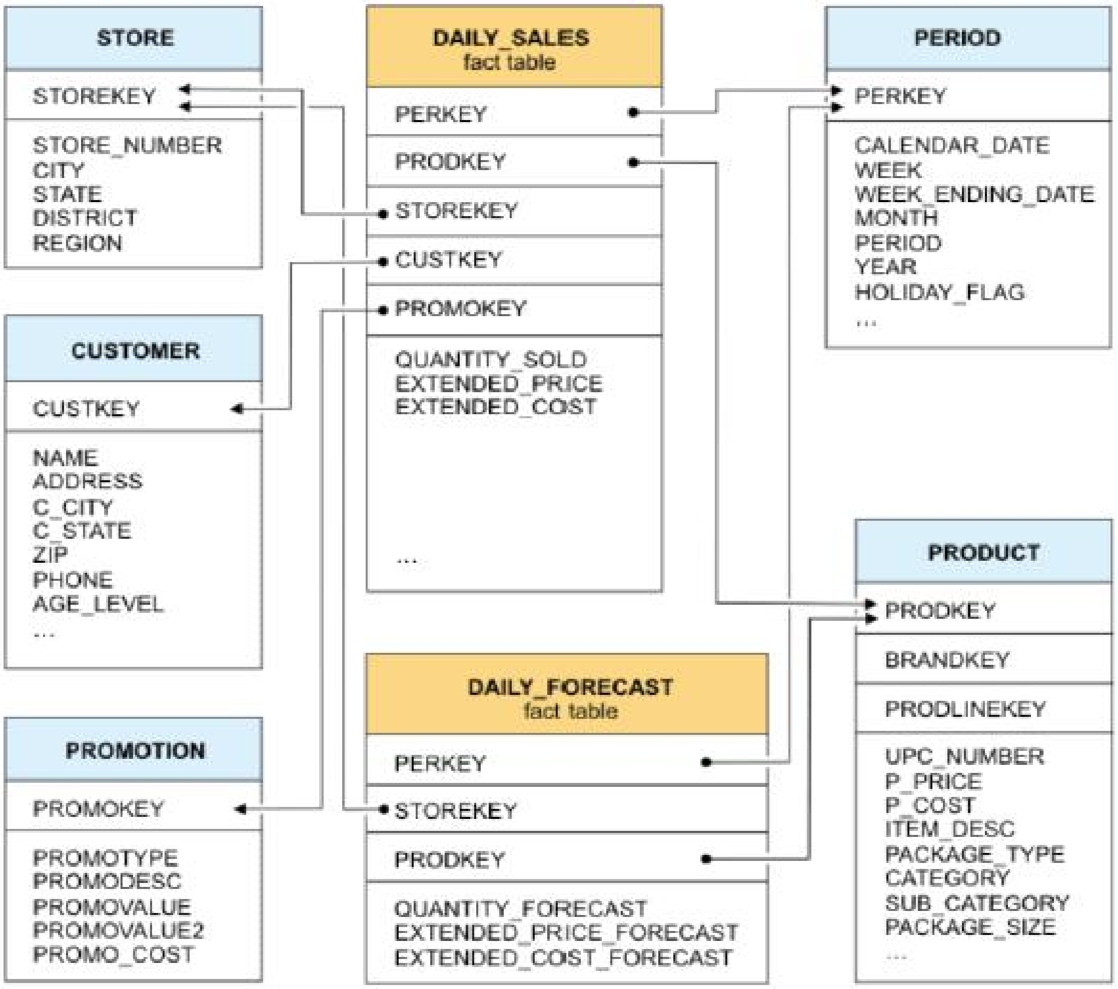


```
select sale.salesamount from
date orderdate,
date shippingdate,
sales sale
where sale.orderdate=orderdate.date and sale.shippingdate=shippingdate.date;
```



==

3. For the given Dimension Model, can you please generate a sql to get the total divergence between Quantity sold and Quantity Forecast for the current month for all the stores:



```
select
      ((SELECT sum(quantity_sold), store_key
        from d.dailysales, period, s.store
        where date = tochar(sysdate,'MM'),
        d.storekey = s.storekey) - (SELECT
        sum(quantityforecast) from dailyforecast df,
        period,    store  s
        where date = tochar(sysdate,'MM'), df.storekey = s.storekey) as divergence;
```

==

4. For the above-mentioned dimension model, please identify the conformed and non-conformed dimensions. Additionally, identify the measure types?

Conformed Dimension : Store, Period, Product.

Non-Conformed Dimension : Customer, promotion.

Quantity sold : Additive Measure.

Quantity_forecast sold : Additive Measure.

Extended_cost : Semi-Additive Measure

Extended_price : Semi-Additive Measure

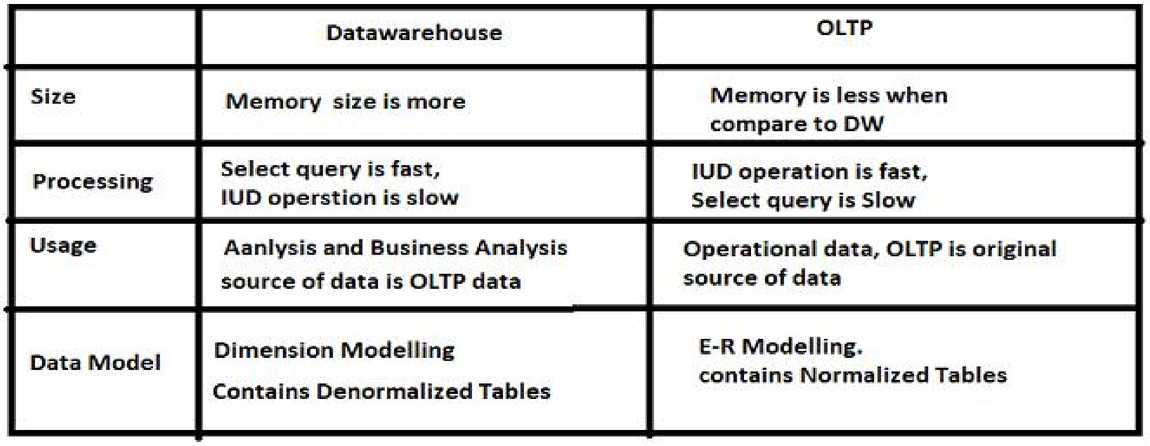
Extended_cost_forecast : Semi-Additive Measure

Extended_price_forecast : Semi-Additive Measure

==

5. Make a list of differences between DW and OLTP based on Size, Usage, Processing and Data Models.

Difference between DW and OLTP,



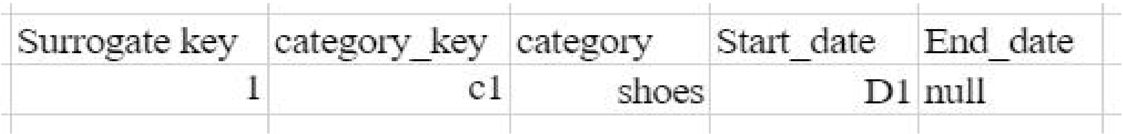
	Datawarehouse	OLTP
Size	Memory size is more	Memory is less when compare to DW
Processing	Select query is fast, IUD operstion is slow	IUD operation is fast, Select query is Slow
Usage	Aanlysis and Business Analysis source of data is OLTP data	Operational data, OLTP is original source of data
Data Model	Dimension Modelling Contains Denormalized Tables	E-R Modelling. contains Normalized Tables

==

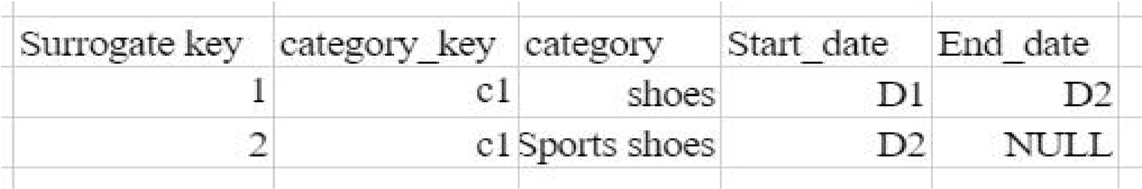
a. Category of a product may change over a period of time. Historical category information (current category as well as all old categories) has to be stored. Which SCD type will be sutiable to implement this requirement? What kind of structure changes are required in a dimenstion table to implement SCD type 2 and type 3.

SCD-2 is the suitable to store the data because its used to store all the historic data but the SCD-3 is only used to store the present data and immediate past data only. As we need to Store all the Historic data we go with SCD-2.

The table changes for SCD-2 are,

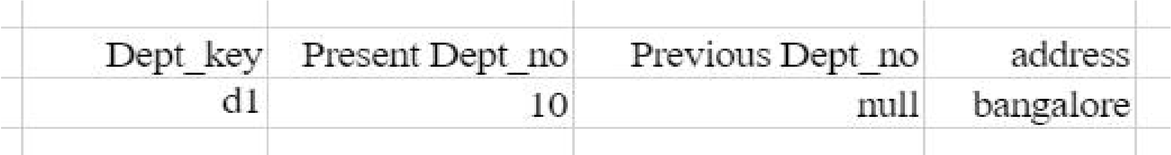


Surrogate key	category_key	category	Start_date	End_date
1	c1	shoes	D1	null

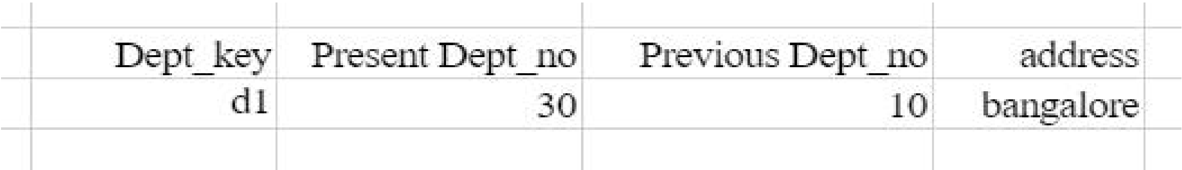


Surrogate key	category_key	category	Start_date	End_date
1	c1	shoes	D1	D2
2	c1	Sports shoes	D2	NULL

The table changes for SCD-3 are,



Dept_key	Present Dept_no	Previous Dept_no	address
d1	10	null	bangalore



Dept_key	Present Dept_no	Previous Dept_no	address
d1	30	10	bangalore

As shown in above the SCD-2 has all the historic data but SCD-3 has only present and immediate past data. so we use SCD-2 to store all category data.

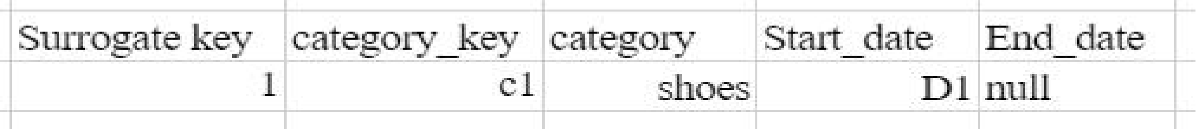
==

b. What is surrogate key? Why it is required?

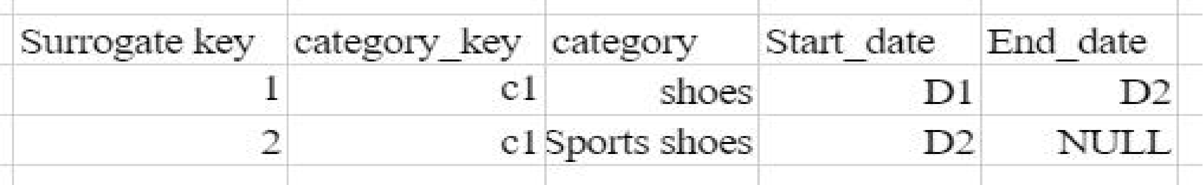
Surrogate key is a numeric key added to SCD-2 table which will be added when ever there is a changes made in table. Surrogate keys are used instead of actual key because is numeric value and easy to map. Surrogate key can have multiple copy of same key with modified data.

They are auto generated key and has no special meaning.

for example,



Surrogate key	category_key	category	Start_date	End_date
1	c1	shoes	D1	null



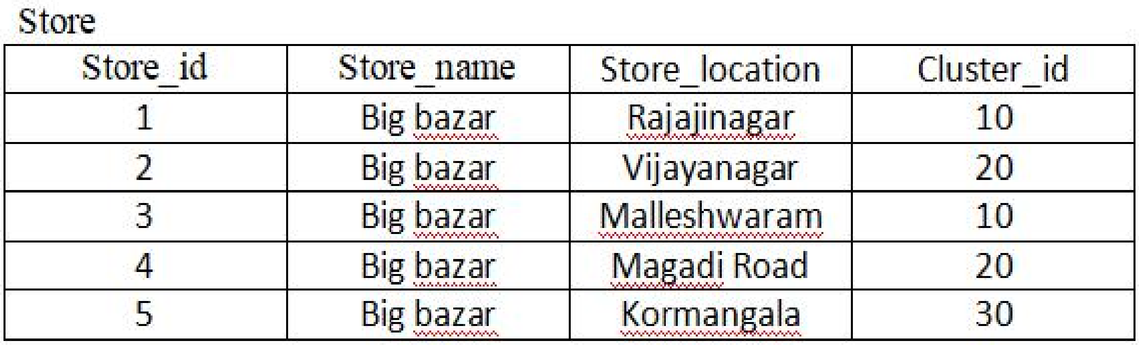
Surrogate key	category_key	category	Start_date	End_date
1	c1	shoes	D1	D2
2	c1	Sports shoes	D2	NULL

In the above table the Surrogate key increases but have same key with the modified data.

==

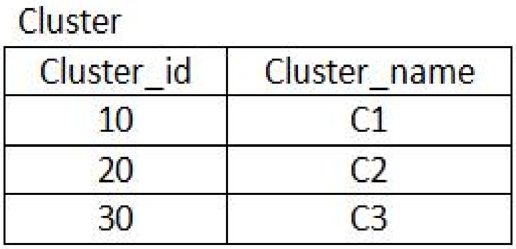
c. Stores are grouped in to multiple clusters. A store can be part of one or more clusters. Design tables to store this store-cluster mapping information.

Consider Big Bazar Shop, they are located in more than one places. This are grouped in the Cluster like C1, C2etc., So the many shop will come under same zone and even might share with other Cluster. Like in the below table we might see Rajajinagar and Malleshwaram



Store

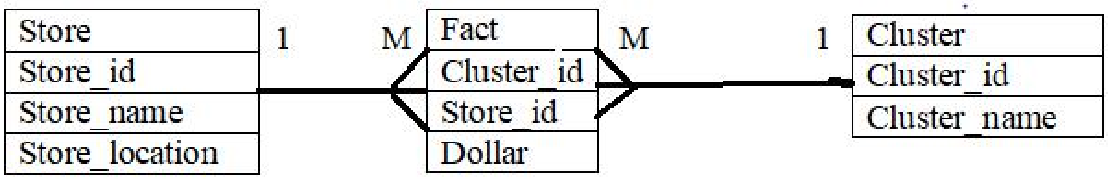
Store_id	Store_name	Store_location	Cluster_id
1	Big bazar	Rajajinagar	10
2	Big bazar	Vijayanagar	20
3	Big bazar	Malleshwaram	10
4	Big bazar	Magadi Road	20
5	Big bazar	Kormangala	30



Cluster

Cluster_id	Cluster_name
10	C1
20	C2
30	C3

The below table has a fact table where it has both store_id and cluster_id so that they can be mapped easily.



==

d. What is a semi-additive measure? Give an example.

Semi-additive measures are the one which cannot be summarized, but this can be used for the some analytical function. for example, Bank account. In bank account the customer can see the individual balance but the manager can see the overall balance of bank by adding all the customers balance.